**SAPI Expanded Checklist**

**Statistical Analysis Plan with Initial data analysis for observational studies**

|  |  |  |
| --- | --- | --- |
| **Section/Topic** | **Item** | **Checklist item** |
| **ADMINISTRATIVE INFORMATION** |
| Project title | 1.1 | **State the title of the analysis project** |
| Project documents | 1.2 | **Provide links to documents describing the analysis project for this SAPI, if available*** For example, links to the research protocol or data management plan including version numbers
 |
| Ethics approval | 1.3 | **Specify if this project requires ethics approval and provide details, if applicable*** If ethics approval is required for this project, indicate the status (not yet submitted, submitted, approved, exempt)
* Name the institutional research board(s) or ethics committee(s) relevant for conducting the analyses
* Provide details of how this analysis project is covered by existing informed consent, if applicable
* Alternatively, refer to the documents where this information is provided
 |
| Names and contact | 1.4 | **Provide names, affiliations, and contacts of key project team members*** List all author(s) of this document
* List all principal investigator(s) of this analysis project
* List all data analyst(s)
* List any oversight committee members, if applicable
 |
| **PROJECT BACKGROUND** |
| Research aims and objectives   | 2.1 | **Describe the research aims of this analysis project (relevant to the statistical analysis)*** Refer to applicable project documents for background and rationale
 |
| Target population  | 2.2 | **Describe the target population of interest for the research objectives*** For example, explain health care context, geographic location, age, sex, gender, specific disease to which the results of the analysis project apply
 |
| **OBSERVATION UNITS** |
| Design | 3.1 | **Describe the design of the research projects, for example cross-sectional, cohort, case-control, longitudinal, surveys, target emulation trial*** Describe methods for defining observation units
* Specify follow-up times, if applicable
* Describe methods for assessment and collection of data,

and corresponding instruments (e.g. questionnaires or laboratory tests), along with their reliability and validity, if known.* Alternatively, refer to the documents where this information is provided
 |
| Data sources  | 3.2 | **Describe the sources of data, for example routinely collected electronic health records, registries, surveys, cohort study, or other types of observational studies.**  * Describe the design of the data source (structure and origin), if applicable
* Include the name and/or DOI of the data source or study, if applicable
* Discuss potential for information biases or selection biases of the data source or study of relevance for the analysis project
* Alternatively, refer to the documents where this information is provided
 |
| Eligibility criteria  | 3.3 | **Specify eligibility criteria for the analysis project based on information available at baseline*** Prepare a list of eligibility criteria ordered for a flow diagram starting with the data source and successive steps for inclusion criteria and then exclusion criteria
* Specify the observation period for which data will be used with start and end dates. This may refer to periods for examinations, diagnoses, visits, events, or follow-up, for example.
* Describe the setting such as health care context or geographic location
 |
| Sampling | 3.4 | **Describe the observations for which data is accessed and observation period when such data is/was collected from observation units*** Describe the sampling strategy, if applicable
* Describe the index date which will be used to synchronize observation periods between observation units in analyses, if applicable. For example, the index date could be the date of diagnosis, the date of treatment assignment, the date at which predictors are measured. Define the index date for each objective.
* Alternatively, provide links to the documents where this information is provided
 |
| Data sets  | 3.5 | **Describe how the data is provided for this analysis project, for example format and content of data sets** **and summarize data preprocessing** * Describe how the data was processed from their raw state after data collection until transfer for use in this analysis project. For example, linking of datasets, deletion of cases and variables, data checking and cleaning.
* Include a data dictionary or refer to documents where this information is available. This includes variable abbreviations, variable names, values, data sources, data standards, or expectations about the data
* Alternatively, refer to the documents and reproducible code where this information is provided, if available
 |
| **VARIABLES** |
| Variables used in the main data analysis | 4.1 | **Define all variable(s) and their roles that will be used in the analyses to answer research questions*** Include how they are measured, units of measurement, and the time point(s) of measurements relative to the index date
* Define roles of variables such as outcome, exposure, mediators, or confounding variables for each objective including DAGs, if applicable
* Include a mapping of variables aligned with each objective, if applicable
 |
| Other variables  | 4.2 | **Define any variables that are not directly used to answer the research objectives, but which provide information about the observation units*** These variables can be used for evaluating data quality, structuring IDA reports, or generating statistical weights. Examples are process variables such as centers, time stamps, or other design variables
 |
| **METHODS: MAIN DATA ANALYSIS (MDA)** |
| Description of observation units | 5.1 | **Describe the methods of analysis to describe the characteristics of the observation units** For example, describe which variables will be included in suchdescriptions, type of summary statistics used, or graphical descriptions. A template for a table summarizing the characteristics can be included here.* Specify if these data descriptions will be stratified by any variable, and by which, if applicable
* Specify any association analyses to be conducted, e.g. by means of correlation analysis, and the type of association measure to be computed or the graphical display of such associations
 |
| Main data analysis methods  | 5.2 | **Describe the methods of analysis for each research objective*** Define the outcome variables for each analysis
* Describe which variables will be included in each analysis and how they are handled, for example regarding transformations, functional forms, or interactions
* Describe reasons for including other variables, for example exposures, to control for confounding, to adjust an estimated association, or because they are known or suspected predictors
* Describe the analysis population(s) for each analysis (if different, define the primary analysis population)
* Define estimands, if applicable
* Specify planned subgroup analyses, if applicable
 |
| Assumptions and diagnostics | 5.3 | **State any statistical assumptions of each analysis. Specify all measures and diagnostics used to evaluate statistical assumptions and appropriateness of analyses, including graphical tools*** Specify all measures used to evaluate model performance, if applicable
* If missing data are expected, describe how this will be handled with details about any imputation or augmentation method
* If measurement errors are expected, describe how they will be addressed. Specify planned sensitivity analyses, if applicable
* Specify planned sensitivity analyses, if applicable
* Describe potential limitations of the analyses
 |
| Sample size | 5.4 | **Describe available information about sample size, number of outcome events, or follow-up time, if applicable*** If the provided numbers are based on expectations provide the rationale
* Justify that the sample size is sufficient for the research objectives, potentially including a sample size or power calculation
* Provide all details and assumptions needed to replicate the sample size or power calculation independently
 |
| Software | 5.5 | **Describe software used for analyses, visualizations, data management, data archiving, or backups** |
| **METHODS: INITIAL DATA ANALYSIS (IDA)** |
| Data preparation | 6.1 | **Describe the methods and processes for preparing data for the analyses** * Describe how compliance of the delivered data with pre-specified structural and technical requirements is assessed (see sections Data and Variables). For example, formats of variables are as expected or ID variables to link data sets are correct
* Describe how variables related to time will be summarized, for example temporal patterns of observations, time differences between observations, or follow-up distribution, if applicable
* Describe how new variables needed for the main data analysis are computed
* Describe any preparation or formatting of data needed for the analyses (multistate or wide and long format)
 |
| Unit missingness | 6.2 | **Describe methods to identify the degree of unit missingness** * Unit missingnessdescribes the lack of data availability from the data source for an observation unit, e.g. if a participant was enrolled in a study but then did not provide any data, or if a participant was invited to participate but did not respond. For some analyses, this could include participants who dropped out before the relevant follow-up point.
* Describe methods to compare characteristics of those with and without unit missingness, where possible. For example, specify which characteristics of participants will be compared to those from census data, or specify what baseline data will be compared for those who did and did not drop out before the study sample was recruited.
 |
| Unit profile | 6.3 | **Describe the methods to summarize the temporal or structural pattern of observations for each observation unit** * Unit profile refers to such patterns of observations, for example the number of units enrolled over time, over geographical locations, or other relevant structures
 |
| Item missingness | 6.4 | **Describe methods to examine item missingness (missing values in variables)*** Describe methods to explore the number and proportion of missing values for each variable and reasons for missingness, if known
* Describe methods to evaluate the number of complete observations for each analysis, if applicable
* Describe methods to examine patterns of missingness and relate these to completely observed variables, for example to outcome and exposure variables or any important structural variables that are completely observed.
 |
| Univariable descriptions  | 6.5 | **Describe methods to summarize the distribution of each variable used in the analyses using numerical or graphical summaries**  |
| Multivariable descriptions  | 6.6 | **Describe any multivariable descriptive statistics and graphical presentation** * Specify (“structural”) variables which will be used to help structure the results of multivariable analyses
* Specify which associations between covariates with these structural variables will be evaluated and the numerical measures or graphical summaries that will be used
* This does not include evaluating associations with outcome variables
 |
| **EVALUATION AND UPDATES** |
| Evaluating the SAPI | 7.1 | **Indicate if an update of the SAPI is needed after IDA*** This information is provided after completion of IDA
* If there is no need for an update, this should be stated
 |
| Updating the SAPI, if applicable | 7.2 | **List all SAPI section(s) affected by updates and state the reasons for the updates** * This may include changes to inclusion or exclusion criteria for observation units or updated information about the variables, for example ranges, categories, transformations
* Updates to the main data analysis plan may also be needed
 |
| **SUPPLEMENT** |
| Key references | 8.1 | **Include references related to statistical methods, research aims, or other background information** |
| Sustainable handling of analysis outputs | 8.2 | **Describe the measures to ensure sustainable handling of analysis outputs** * List the analysis documents, for example final SAPI, IDA report, MDA report, data quality report, code files, updated data files
* Describe data formats and documentation that will be used to ensure interoperability and reuse by the research team or third parties
* Indicate where and how these outputs will be archived or who they will be shared with
 |